Automatic identification methods on a corpus of twenty five fine-grained Arabic dialects

Salima Harrat, Karima Meftouh, Karima Abidi, Kamel Smaïli

Automatic identification methods on a corpus of twenty five fine-grained Arabic dialects

Salima Harrat[1], Karima Meftouh[2], Karima Abidi[3], and Kamel Smaïli[3]

[1] École Normale Supérieure de Bouzaréah
Algiers - Algeria
slmhrrt@gmail.com
[2] Badji Mokhtar University
Annaba - Algeria
karima.meftouh@univ-annaba.dz
[3] Loria - Univ. Lorraine
Nancy - France
{karima.abidi,kamel.smaili}@loria.fr

Abstract. This research deals with Arabic dialect identification, a challenging issue related to Arabic NLP. Indeed, the increasing use of Arabic dialects in a written form especially in social media generates new needs in the area of Arabic dialect processing. For discriminating between dialects in a multi-dialect context, we use different approaches based on machine learning techniques. To this end, we explored several methods. We used a classification method based on symmetric Kullback-Leibler, and we experimented classical classification methods such as Naive Bayes Classifiers and more sophisticated methods like Word2Vec and Long Short-Term Memory neural network. We tested our approaches on a large database of 25 Arabic dialects in addition to MSA.

Keywords: Arabic dialects · Automatic dialect identification · Dialect resources · Parallel dialectal corpora.

1 Introduction

Standard Arabic is the official language of Arab countries, it is used in formal speech, education, and newspapers. In contrast people, all over the Arab world use Arabic dialects in their everyday conversations. Indeed, Arabic dialects are a variant of the Arabic language (besides Modern Standard Arabic and classical Arabic). Most research classifies Arabic dialects according to East-west dichotomy [8]: Maghrebi dialects (Algeria, Morocco, Tunisia, Lybia, and Mauritania) and middle-east dialects (Egypt, Sudan, Gulf countries and Levantine countries). Another research [25] classifies them according to the ethnic and social diversity of Arab speakers as rural and Bedouin variants.

Arabic dialects differ widely between and within Arab countries. Arabic dialects share a lot of features with standard Arabic which makes them close to each other but also have specific characteristics related to each one. Social

media and mobile telephony have contributed to the increased use of Arabic dialects in a written form. In this context, discriminating between dialects in a multi-dialectal corpus of texts is a challenging issue, especially when dialects belong to regions from the same Arabic country. In this case, it is fine-grained identification where we have to distinguish between very close dialects.

In this paper, we deal with the dialect identification at the sentence level. We used several approaches and experimented different features. The features are those parameters that are supposed to characterize specifically each language. Consequently, they are crucial and not easy to determine.
The remainder of this article is organized as follows: in Section 2 we highlight the most challenges issues related to Arabic dialects identification, we present the most important points that make this task a hard one. Section 3 summarizes relevant research efforts in dialect identification, while Section 4 presents our contribution in this area by describing the four approaches we explored. In Section 5 we give a brief description of the dialectal corpus we used for training our classifiers and Section 6 is allocated to the results of our experiments. Section 7 concludes this paper.

2 Arabic dialects identification challenges

In their oral form, Arabic dialects are relatively easy to distinguish. In fact, prosody and tone bring important information about them. But, in their written form, and compared to other languages, Arabic dialects are difficult to identify. They are similar languages that share a lot of features and words although they may differ from one Arab country to another and from one city to another within the same country. In the following, we enumerate the reasons that make difficult the issue of the identification of Arabic dialects.

- They share a lot of lexical units with modern standard Arabic. Consequently, distinguishing between Arabic dialects is a hard task.
- Some words are shared among Arabic dialects but with different meanings. For example, the Egyptian word ليه which means *why* exists in other dialects like Algerian but with another meaning: *for him*.
- In the conversation, Arab people tend to switch to standard Arabic especially when discussing matters relating to religion. Thus, the use of standard Arabic makes the identification task confusing.
- The lack of dialectal resources such as monolingual and multilingual corpora makes the identification task a challenging issue. Indeed, the identification data-driven approaches require important amounts of data to reach acceptable accuracy rates, such resources are not available for most Arabic dialects.

3 Related work

Several studies in the area of Arabic NLP attempted to deal with the dialect identification issue. Different approaches have been adopted. Early work in this

area[26] used language modeling (LM) based approach to identify the dialect at the sentence level. The authors created for the purpose of this research the Arabic Online Commentary Dataset (OAC) (a collection of 52M-word monolingual dataset rich in dialectal content and annotated thanks to the crowdsourcing principle). Each dialect of this corpus was modeled by a 3-gram LM, then the sentence perplexity was computed to score each sentence of the test corpus.

The same authors in [27] used the previously created Arabic Online Commentary Dataset (with the annotated data) to train classifiers using word and character language models. They use 1-gram, 2-gram of words and 1-gram, 3-gram and 5-gram of letters. They conducted two-way classification: MSA vs. Dialect, and multi-way classification: (MSA, multiple dialects). They explored two identification approaches, by creating: a first system where they use MSA-only data and attempt to determine how MSA-like a sentence is. They extracted a vocabulary of 2.9M of words from the Arabic Gigaword Corpus. Then each sentence is given an OOV percentage of dialectal words, when this percentage reached a fixed threshold, the sentence is considered as being dialectal. The second system used perplexity to classify sentences, a language model using only MSA data was trained on 43M words extracted from the article bodies of the AOC. When exceeding a perplexity threshold the sentence is classified as being dialectal. The authors conclude that classifiers trained with dialectal data (with word 1-gram LM) significantly outperform classifiers which use MSA data only.

Later supervised approach was used to address dialect identification. The authors in [6] proposed a supervised approach to predict whether the sentence is MSA-like or Egyptian. To this end, they trained a Naive Bayes Classifier (NBC) using token based features and perplexity based features, in addition to other features like (percentage of punctuation, numbers, special-characters, number of words & average word-length, etc.). They evaluated their system on the Egyptian part of the OAC described above. In [19], The authors used Markov character-based n-grams language models and NBC trained on social media data for Arabic dialect identification task. They first experimented with 1-gram, 2-gram and 3-gram character-based LMs. Then, they trained NB classifiers using the three LMs as features. The identification task covered 18 Arabic dialects. They also conducted experiments on 6 groups of dialects defined regards to geographical repartition. The achieved results show that NB classifier outperforms the character-based n-gram Markov model for most Arabic dialects. In the same vein, the best accuracy rates are got with NBC with 2-gram LM features.

The authors of [20] dealt with fine-grained dialect identification. They attempted to identify 25 dialects of different Arabic cities in addition to MSA. They also perform dialect identification within 6 geographical regions. They used a Multinomial Naive Bayes (MNB) classifier for the learning task. The classifiers are trained by word and character n-gram LMs. They conduct a set of experiments by varying the use of features from character/word 1-gram to 5-grams and by combining them. The best accuracy was reached with features

from word 1-gram LM, 1-gram to 3-gram character LM and Character/Word 5-gram LM probability scores.

Other research used SVM approach to address the dialect identification issue. In [3], the authors presented a multi-dialect, multi-genre, human annotated corpus of dialectal Arabic (Egyptian, Gulf, Levantine, Maghrebi, and Iraqi) extracted from online newspaper commentary and Twitter. They used crowdsourcing via mechanical Turk to annotate the data. In terms of size, the corpus contains 27239 newspaper comments including 583K words and 40229 tweets including 666K words. With these data, they dealt with dialect identification by combining LMs and machine learning. They use two classifiers: SVM with a linear kernel and NB classifiers both trained on word n-gram LM features. The results show that the 1-gram based model performs better than 2-gram/3-gram based models for both SVM and NBC. Moreover, the NB classifier gives better results.

A similar method was used in [12] where the authors used the Multidialectal Parallel Corpus of Arabic [2] to perform dialect identification. They used a SVM classifier with word 1-gram/2-gram LMs and character 1-gram to 4-gnam LMs features (without any preprocessing step). The authors used SVM to perform multi-class classification. They also used a meta-classifier (SVM based) trained by the class probability outputs of lower classifiers (described above). Each lower (SVM) classifier is learned from one feature type. The authors reported an accuracy of 74% on the 6-way identification task. For 2-way identification, the accuracy reached 94% and The best features are those related to 3gram.

The authors of [4] used lexical, phonological, morphological, and syntactic features to distinguish between dialectal Egyptian and MSA. They used Random Forest (RF) classification for two-way dialect-MSA identification. The RF classifier was trained on the Egyptian side of OAC [26] and 150K MSA sentences from an English-MSA parallel corpus. It used word 1-gra/2-gram/3-gram LMs and character 1-gram to 5-gram LMs as features. The authors show that the RF classifier performs better when it uses features extracted from segmented data in addition to lexical features.

Another interesting work is that described in [5]. It presents Aida2, a token and sentence level dialect identification system that distinguishes between MSA and Egyptian dialect. It uses a set of classifiers to deal with the identification task on the two levels. At token level, the identification is considered as sequence labeling task. The authors used Conditional Random Field (CRF) classifier which is trained by using decisions from several underlying components: MADAMIRA morphological analyzer [17], a tokenized 5-gram Language Model, a compiled lexicon of Arabic modality triggers, and a Named Entity Recognizer. The output of this first module is then given to the sentence level identification module which relies on two independent underlying classifiers. The first one uses tokenized-level LMs, thus it yields detailed and specific information about the tokens. The second one is based on surface forms MSA and Egyptian dialect 5-gram LMs. Each of the two classifiers gives a class label and a confidence score

to the input sentence. Given this information, a Decision Tree classifier provides the final class of the sentence.

In the same vein, the authors of [21] dealt with the identification of code-switching between MSA and Moroccan dialect in discussion boards and blog text. The identification task is considered as a sequence labeling problem which the authors treat by using CRF. Regards to the data, the authors created their annotated corpus from scratch by downloading discussion boards and blogs and proceeded to the annotation for the purpose of identification. To train the CRF classifier, they used 5 types of features like the words and their surrounding words with their affixes, structural properties such as if the word contains numbers, character language models and lexical knowledge from an external source such as word lists. The authors combined these features in order to identify the best combination which gives the best accuracy.

4 Identification approaches

In the following, we present the different approaches we tested and evaluated.

4.1 Long Short-Term Memory neural network approach

A recurrent neural network (RNN) in which the connections are made between units which form a directed cycle, which allows it to exhibit a dynamic temporal behavior for the model. Long Short Term Memory Networks (LSTM)[9] are a special class of neural networks able to learn long-term dependencies. They are designed especially to avoid the long-term dependency problem. Their main characteristic is that they remember information for long periods of time. This class of neural network has been efficient for many NLP tasks such as language modeling [24], sentiment analysis [16], word embedding learning [11], as well as in other area like automatic speech recognition [7] and image captioning [13].

We consider the dialect identification task as a multi-class classification problem that we attempt to solve with Long Short-Term Memory (LSTM) networks: given a sentence s_j, dialect features vectors V_i with their corresponding labels l_i, we have to predict l_j by using V and s_j. We designed a recurrent network classifier that takes as input a vector of characters/words n-grams (for characters n varies from 1 to 5 and for words it varies between 1 and 2). It goes through a LSTM layer, then to a drop out layer to prevent over-fitting. The last layer of the network is a softmax that gives a probability distribution over the different dialect labels.

After several setup configurations, we retained the following parameters for our neural network architecture:

- Input vector dimension is variable, it depends on the vectorization parameters. We used character and word level vectorization with different orders.
- LSTM layer units: 128
- Droupout rate: 0.2

4.2 Word embedding based approach

The idea is to investigate to what extent the semantic information encoded by word embedding can be used to identify the varieties of Arabic dialects. For this reason, we used the CBOW method of Word2Vec model [15] to extract the vector representation of the words. Given the limited size of the dialectal corpora and knowing that neuronal network methods necessitate an important amount material for training, we decided to increase the data by using the infra-lexical information of the provided corpus. That is why each sentence of each dialect of the multi-dialect corpus is segmented into 2, 3, 4 and 5 grams of characters. In addition, the original sentence is kept in the corpus necessary for the training. After this step, only the vectors representing the typical words of each dialect are kept for the test. The typical words are those words or infra-lexical units that are characteristic of a dialect. To identify these units, we kept for a dialect only the units that do not occur in other dialects.

To label a sentence s with its appropriate tag t from the $|D|$ dialects, we calculate the similarity between the units of s and the list of typical words of each dialect as follows.

$$d_k = \frac{1}{|s|} \sum_{i=1}^{|s|} \underset{1 \leq j \leq |L_k|}{min} E(s_i, w_j^k) \tag{1}$$

$$i_l = \underset{1 \leq k \leq |D|}{argmin}(d_k) \tag{2}$$

Where:

- $|s|$ is the number of words of s,
- L_k is the list of typical words of the dialect k,
- E is the Euclidean distance,
- w_j^k is the word j belonging to the list of typical words of the dialect k,
- and $|D|$ is the number of dialects/language (distinct labels).

Then we assign the label l corresponding to the dialect that gives the smallest distance.

4.3 Symmetric Kullback-Leibler for classification

In this approach, we constitute a General Vocabulary (GV) from the different training corpora. The vocabulary is composed of all the words, the bi-grams and with all the infra-lexical units from one to five. Then, the distribution of each dialect is calculated in accordance to GV. Each dialect d_i is then represented by a vector where each dimension is given by $P(u_k|d_i)$. Where u_k indicates a unit of GV and d_i corresponds to the dialect i. All the probabilities are smoothed to avoid zero probabilities for unknown words of the test corpus.
For the test, each sentence is segmented similarly to what has been done for the training. Then we calculate the symmetric Kullback-Leibler measure [10](see

equation 3), we used several years ago to identify emails [1], between the distribution of the test sentence and the distribution of each dialect. We assign then the sentence to the dialect that provides the smallest score.

$$D(P||Q) = \sum_{x}((P(x) - Q(x))Log\frac{P(x)}{Q(x)} \tag{3}$$

4.4 Multinomial Naïve Bayes (MNB) approach

Naïve Bayes classifiers are widely used in different applications in natural language processing and particularly in text classification[14][18][23] due to their efficiency and their acceptable predictive performance. That is why we consider them to deal with the dialect identification issue . MNB estimates the conditional probability of a particular term given a class as the relative frequency of the term t in all documents belonging to the class C.

In order to train our MNB classifier, we used 1-gram, 2-gram and 3-gram as features supported by a TF-IDF vector We also used a special character to mark the start of the sentences. We note that we utilized Term Frequency-Inverse Document Frequency (TF-IDF) scores [22].

5 Data description

For training and testing our classifiers, we used the MADAR shared task data [20]. It consists of two parallel multi-dialect corpora:

- The first corpus (MADAR-Corpus26) is composed of parallel sentences translated to 25 dialects of several cities from the Arab countries (see Table 1), in addition to modern standard Arabic. Each dialect/language includes 1600 sentences for training and 200 sentences for test purpose.
- The second corpus (MADAR-Corpus6) is a collection of 10K additional sentences translated to the dialects of five selected cities: Beirut, Cairo, Doha, Tunis, and Rabat.

In Table 2, we give an example of parallel sentences from MADAR-Corpus26 (the first corpus).

Table 1. MADAR-Corpus26 Countries and cities.

Country	City		Country	City	
Algeria	Algiers	ALG	Palestine	Jerusalem	JER
Morocco	Rabat	RAB	Syria	Beirut	BEI
	Fes	FES		Damascus	DAM
Tunisia	Tunis	TUN		Aleppo	ALE
	Sfax	SFX	Iraq	Mosul	MOS
Libya	Tripoli	TRI		Baghdad	BAG
	Benghazi	BEN		Basra	BAS
Egypt	Cairo	CAI	Saudi Arabia	Riyadh	RIY
	Alexandria	ALX		Jeddah	JED
	Aswan	ASW	Oman	Muscat	MUS
Sudan	Khartoum	KHA	Qatar	Doha	DOHA
Jordan	Amman	AMM	Yemen	Sanaa	SAN
	Salt	SAL			

6 Experiments

We built a set of classifiers based on the approaches described above by using the two MADAR corpora (MADAR-Corpus26 and MADAR-corpus6). For each classifier, we tested several combinations of features to identify the ones that increase the accuracy values. We report in Table 3 the best-achieved results and in Table 7 the features that yield the best accuracy rate for each approach.

The best achieved results are those got with the multinomial NB approach, followed by the LSTM, then Kullback-Leibler, while the word embedding values come last. Sophisticated approaches did not give the intended results. We expected to have better or at least equivalent results with the neural network approach. But the experiments show that MNB performs better. This is due in our opinion to the size of the training data; Indeed neural networks require an important amount of data to perform best.

In addition, 6-way identification classifiers perform better than 26-way identification. This is a natural and expected result since the confusion is reduced when using fewer dialects and more data. It is worth noting 6-way identification results follows the same scale of values as 26-way identification, MNB results remain the best followed by LSTM, Kullback-Leibler and W2Vec values. But we can mention that the results of the LSTM and the symmetric Kullback-Leibler are close to each other.

For the MNB classifiers (for convenience we refer to them by MNB-MADAR-Corpus26 & MNB-MADAR-Corpus6) which achieved the best scores, we computed respectively, Precision, Recall, and F1-score at class level (see Table 4 and Table 5). We also generated the confusion matrix of these classifiers in order to have an idea about the dialects they recognize better than others and the errors they make. For presentational reasons, we report in Table 6 a summary of MNB-MADAR-Corpus26 confusion matrix, while in Fig. 1 we show the confusion matrix of MNB-MADAR-Corpus6.

Table 2. Example of parallel sentences from MADAR-Corpus26.

City	Dialect or language	Sentence
	MSA	هناك ، أمام بيانات السائح تماما .
Beirut	BEI	صار هونيك ، بالظبط قدام مكتب استعلامات السياح .
Cairo	CAI	ده قدامك هناك ، يادوبك قدام مكتب استعلامات السياحة .
Doha	DOH	هو ذاك الصوب ، بالضبط جدام استعلامات السياح بالضبط .
Rabat	RAB	راه تما ، مقابل مكتب استعلامات السياح بالضبط .
Tunis	TUN	اهوكا غادي ، بالضبط قدام البيرو متاع الارشادات السياحية .
Alexandria	ALX	هو هناك ، قدام الاستعلامات السياحية على طول .
Algiers	ALG	راهو لهيك ، بالضبط قدام المكتب تع معلومات السياح .
Aswan	ASW	هناك ، قدام مكتب ارشادات السياح على طول .
Damascus	DAM	موجود هنيك ، قدام مكتب معلومات السياح بالزبط .
Jeddah	JED	شوفه هناك ، قدام مكتب المعلومات السياحية بالضبط .
Ryadh	RIY	هناك ، بالضبط مقابل مكتب معلومات السياح .
Sfax	SFX	أوكي غادي ، قدام مكتب الإرشادات السياحية بالضبط .
Baghdad	BAG	موجود هناك ، بالضبط مقابيل مكتب المعلومات السياحية .
Meaning	There, just in front of tourist information	

Table 3. Dialect identification results using different approaches.

Training Corpus	MADAR-Corpus 26			MADAR-Corpus 6		
Approach	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Word Embedding	50.11	49.90	49.74	83.96	83.90	83.83
Symmetric Kullback-Leibler	53.21	68.27	53.79	89.05	89.48	89.03
Multinomial Naïve Bayes	69.80	69.15	**69.09**	92.54	92.50	**92.50**
LSTM networks	58.04	61.54	58.33	89.23	89.17	89.18

For 26-way classification, the dialects with low confusion rates were better identified than others. The Mosul dialect (MOS) achieved the best scores. Although it is an Iraqi dialect, it is well distinguished compared to the other Iraqi dialects (BAG and BAS). These two last are confused by a rate of 18.5%. Similarly, the classifier tends to confuse the dialects belonging to the same countries. The most confused dialect pairs are RAB & FES, SFX & TUN, CAI & ASW, ALX & ASW and BEN & TRI, in addition to MUS which is the most confused dialect with MSA. Furthermore, the Levantine dialects because of their closeness are also confused with each other (AMM & JER and DAM & AMM). The recall values of all these dialects are lower compared to other values recorded for dialects such as ALG and SAN that are the only ones belonging to Algeria and Yemen in this order.

Table 4. MNB-MADAR-Corpus26 Dialect identification results by dialect/language.

Dial./Lang.	Precision	Recall	F1-score	Dial./Lang.	Precision	Recall	F1-score
MOS	83.41	85.50	84.44	ALE	78.12	62.50	69.44
ALG	78.08	85.50	81.62	DOH	72.04	67.00	69.43
SAN	87.79	75.50	81.18	KHA	63.29	75.00	68.65
MSA	71.49	89.00	79.29	BAS	67.15	69.50	68.30
ALX	76.17	81.50	78.74	JED	68.45	64.00	66.15
TRI	69.26	80.00	74.25	CAI	73.97	54.00	62.43
RAB	78.98	69.50	73.94	ASW	59.55	65.50	62.38
FES	72.25	75.50	73.84	RIY	57.14	64.00	60.38
SFX	67.52	79.00	72.81	SAL	61.90	58.50	60.15
BEI	78.70	66.50	72.09	DAM	56.02	60.50	58.17
BEN	70.87	73.00	71.92	JER	54.63	62.00	58.08
TUN	75.14	65.00	69.71	MUS	65.52	47.50	55.07
BAG	76.97	63.50	69.59	AMM	50.43	59.00	54.38

For 6-way classification, the scores are better. The most confused dialects are RAB & TUN followed by CAI & DOH, then BEI & DOH and BEI & CAI (with the same confusion rate), while the most confused dialects with MSA are DOH and CAI.

Table 5. MNB-MADAR-Corpus6 Dialect identification results by dialect/language.

Dialect/Language	Precision	Recall	F1-score
MSA	95.09	96.80	95.94
RAB	94.04	93.10	93.57
TUN	94.25	91.80	93.01
BEI	93.03	90.70	91.85
DOH	88.21	92.80	90.45
CAI	90.72	89.90	90.31

Table 6. MNB-MADAR-Corpus26 Confusion matrix summary.

Dial./Lang	Recal	Most confused	Confusion %	Less confused[1]
ALE	62.5	DAM/JER	7.5	ALG BAG FES MOS RAB
ALG	85.5	MSA	2.5	ASW BEI BEN CAI JED
ALX	81.5	ASW	9.0	TRI
AMM	59.0	JER	12.5	ALG ALX BEN FES MSA MUS RAB SFX
ASW	65.5	ALX	12.5	DOH JED JER KHA SFX TUN
BAG	63.5	BAS	18.5	ALX ASW BEN CAI DAM JED JER KHA SFX TUN
BAS	69.5	BAG	10.5	ALG ASW BEI DAM JER KHA MUS RAB SAL SAN
BEI	66.5	DAM	6.5	BAS ALX ASW KHA MSA SAN TRI
BEN	73.0	TRI	6.5	ALX JED MOS RAB
CAI	54.0	ASW	17.5	BAS ALE MOS RAB
DAM	60.5	AMM	11.5	ALX ASW BAG BAS MUS SFX TUN
DOH	67.0	RIY	6.0	ALE JER MSA TRI
FES	75.5	RAB	11.5	ALE AMM ASW BEN KHA MUS RIY TUN
JED	64.0	RIY	6.5	ALX MSA
JER	62.0	AMM	9.5	ASW BAG CAI MUS RIY SAN
KHA	75.0	MSA	2.5	ALE BAS DOH SAN SFX TUN
MOS	85.5	BAS	4.0	ALG BEN DAM FES KHA MSA MUS RIY SAL TUN
MSA	89.00	MUS	2.5	ALE BAG BAS BEN JED MOS MOS SFX TRI
MUS	47.5	MSA	15.0	ALX ALE BEI JER RAB SAL SAN TRI RAB SAL SAN TRI
RAB	69.5	FES	18.5	BEN DOH JER MUS TUN
RIY	64.0	MUS	5.5	AMM CAI JER TRI
SAL	58.5	AMM /JER	9.0	BAG DOH FES MOS MUS SAN SFX
SAN	75.5	RIY	4.0	ALE ALG ASW JER MSA RAB
SFX	79.0	TUN	10.5	ASW BAG BEI JED JER KHA MSA MUS RAB SAL SAN
TRI	80.0	BEN	8.0	ALG AMM BAG BAS CAI DAM MOS RAB SAN
TUN	65.0	SFX	19.0	ALE ASW FES JED JER

[1] Confusion rate is equal to 0.5 for all these classes.

In terms of features, we confirm that using n-grams features helps to increase accuracy. All the classifiers perform better when they are fit with such information. Character n-grams order varies from 1 to 5, while for word n-grams lower order (1 and 2) achieve the best results. It should be noted that for the MNB classifier, we used sentence likelihood computed from the 26 word uni-gram language models.

Table 7. The dialect features used in the different approaches.

Approach	Word n-grams features	Character n-grams features
Word Embedding		2-gram to 5-gram
Symmetric Kullback-Leibler	1-gram and 2-gram	1-gram to 5-gram
Multinomial Naïve Bayes	1-gram to 2-gram	1-gram to 5-gram +LMs Prob
LSTM networks	1-gram	4-gram

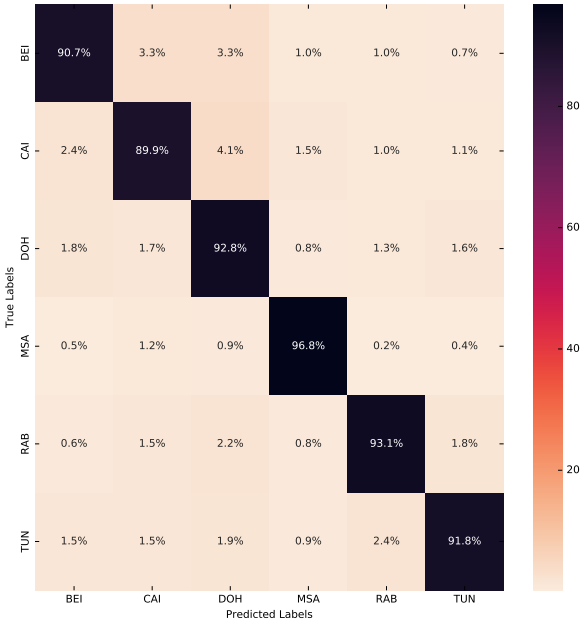

Fig. 1. MNB-MADAR-Corpus6 confusion matrix.

7 Conclusion

In this paper, we explored several approaches to tackle the issue of dialect identification with a set of 25 dialects belonging to some cities from the Arab countries in addition to MSA. We considered neural network approaches by using words embedding and LSTM networks. Unfortunately, the achieved results were not as what we expected, the size of the available training data was not sufficient to learn such classifiers. For W2Vec approach, we get the worst results (F1-score of 49.90 vs 61.54 from LSTM method). In the same vein, we experimented with the symmetric Kullback-Leibler distance. The obtained results did not exceed F1-score of 53.79 but with a recall of 68,27. The best results were achieved by the Multinomial Naïve Bayes classifier. It performs better than all other classifiers with an F1-score of 69.09. All the described classifiers were trained by using different features combinations. The character and word n-grams remain the best features for text classification, especially of Arabic dialects.

References

1. Bigi, B., Brun, A., Haton, J.P., Smaïli, K., Zitouni, I.: A comparative study of Topic Identification on Newspaper and E-mail. In: Proceedings of the 8th International Symposium on String Processing and Information Retrieval - SPIRE'01. pp. 238–241. Laguna de San Rafael, Chili (2001)
2. Bouamor, H., Habash, N., Oflazer, K.: A Multidialectal Parallel Corpus of Arabic. In: Proceedings of the Language Resources and Evaluation Conference, LREC-2014. pp. 1240–1245 (2014)
3. Cotterell, R., Callison-Burch, C.: A multi-dialect, multi-genre corpus of informal written arabic. In: LREC. pp. 241–245 (2014)
4. Darwish, K., Sajjad, H., Mubarak, H.: Verifiably effective arabic dialect identification. In: EMNLP. pp. 1465–1468 (2014)
5. Elfardy, H., Al-Badrashiny, M., Diab, M.: Aida: Identifying code switching in informal arabic text. EMNLP p. 94 (2014)
6. Elfardy, H., Diab, M.: Sentence Level Dialect Identification in Arabic. In: ACL (2). pp. 456–461 (2013)
7. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. IEEE (2013)
8. Hetzron, R.: The Semitic Languages. Routledge language family descriptions, Routledge (1997), https://books.google.dz/books?id=nbUOAAAAQAAJ
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)
11. Li, J., Lin, X., Rui, X., Rui, Y., Tao, D.: A distributed approach toward discriminative distance metric learning. IEEE transactions on neural networks and learning systems **26**(9), 2111–2122 (2014)
12. Malmasi, S., Refaee, E., Dras, M.: Arabic dialect identification using a parallel multidialectal corpus. In: International Conference of the Pacific Association for Computational Linguistics. pp. 35–53. Springer (2015)
13. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)
14. McCallum, A., Nigam, K., et al.: A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. vol. 752, pp. 41–48. Citeseer (1998)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (Workshop) (2013), http://arxiv.org/abs/1301.3781
16. Pal, S., Ghosh, S., Nag, A.: Sentiment analysis in the light of lstm recurrent neural networks. Int. J. Synth. Emot. **9**(1), 33–39 (2018). https://doi.org/10.4018/IJSE.2018010103, https://doi.org/10.4018/IJSE.2018010103
17. Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.M.: Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland (2014)

18. Rish, I., et al.: An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. vol. 3, pp. 41–46 (2001)
19. Sadat, F., Kazemi, F., Farzindar, A.: Automatic identification of arabic dialects in social media. In: Proceedings of the first international workshop on Social media retrieval and analysis. pp. 35–40. ACM (2014)
20. Salameh, M., Bouamor, H.: Fine-grained arabic dialect identification. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1332–1344. Association for Computational Linguistics (2018), http://aclweb.org/anthology/C18-1113
21. Samih, Y., Maier, W.: Detecting code-switching in moroccan arabic social media. SocialNLP@ IJCAI-2016, New York (2016)
22. Spärck Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation **28**, 11–21 (1972)
23. Su, J., Shirab, J.S., Matwin, S.: Large scale text classification using semi-supervised multinomial naive bayes. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 97–104. Citeseer (2011)
24. Sundermeyer, M., Schlüter, R., Ney, H.: Lstm neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association (2012)
25. Watson, J.C.: Phonology and Morphology of Arabic. Phonology of the World's Languages, Oxford University Press, New York (2007)
26. Zaidan, O.F., Callison-Burch, C.: The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 37–41. Association for Computational Linguistics (2011)
27. Zaidan, O.F., Callison-Burch, C.: Arabic dialect identification. Computational Linguistics **1**(1) (2012)